

Optimal Design in Geostatistics under Preferential Sampling*

Gustavo da Silva Ferreira[†] and Dani Gamerman[‡]

Abstract. This paper analyses the effect of preferential sampling in Geostatistics when the choice of new sampling locations is the main interest of the researcher. A Bayesian criterion based on maximizing utility functions is used. Simulated studies are presented and highlight the strong influence of preferential sampling in the decisions. The computational complexity is faced by treating the new local sampling locations as a model parameter and the optimal choice is then made by analysing its posterior distribution. Finally, an application is presented using rainfall data collected during spring in Rio de Janeiro. The results showed that the optimal design is substantially changed under preferential sampling effects. Furthermore, it was possible to identify other interesting aspects related to preferential sampling effects in estimation and prediction in Geostatistics.

Keywords: optimal design, Geostatistics, preferential sampling, point process.

1 Introduction

Diggle et al. (2010) presented a novel methodology to perform inference in the traditional Geostatistical model under preferential sampling. They assumed that the sample design could be described by a log-Gaussian Cox process (Møller et al., 1998) and performed maximum likelihood estimation for the model parameters through simulation. In addition, they have made simulations to evaluate the effect of preferential sampling on parameter estimation in Geostatistics, concluding that this was not negligible. In all simulations, after obtaining unbiased estimates about the model parameters, the spatial prediction was made by the *plug-in* method, according to the classical approach to perform inference in Geostatistics (see Cressie, 1993; Diggle et al., 1998).

Based on this work, many others questions have emerged, and one of them was related to the influence of preferential sampling in the optimal design choice. This procedure is widespread in Geostatistics literature. There are several recent papers dealing with this, such as Zidek et al. (2000); Fernández et al. (2005); Zhu and Stein (2005); Diggle and Lophaven (2006); Gumprecht et al. (2009); Boukouvalas et al. (2009); Müller and Stehlík (2010), among others. The advances made by Müller (1999); Müller et al. (2004) and Müller et al. (2007) that propose methods based on maximization of utility functions are especially relevant. Since these procedures incorporate the Bayesian

*Related articles: DOI: [10.1214/15-BA944A](https://doi.org/10.1214/15-BA944A), DOI: [10.1214/15-BA944B](https://doi.org/10.1214/15-BA944B), DOI: [10.1214/15-BA944C](https://doi.org/10.1214/15-BA944C); rejoinder at DOI: [10.1214/15-BA944REJ](https://doi.org/10.1214/15-BA944REJ).

[†]National School of Statistical Sciences, Brazilian Institute of Geography and Statistics, Rio de Janeiro, Brazil, gustavo.ferreira@ibge.gov.br

[‡]Department of Statistical Methods, Mathematical Institute, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil, dani@im.ufrj.br

approach and intensive simulation methods in a natural way, they became quite appropriate to perform the optimal design choice in Geostatistics.

Here the uncertainties about model parameters will be considered allowing us to evaluate their impact on spatial prediction (or kriging) under preferential sampling. With this aim, spatial prediction about the underlying process are performed by analysing the predictive distribution conditional on the observed data and the design sample. Then, comparisons between these predictions and those obtained according to the classical approach are made.

However, the most important contribution of this paper is its analysis of preferential sampling effects in the process of obtaining the optimal design in Geostatistical models. Using an approach based on maximizing utility functions (Müller, 1999) to obtain optimal design, the influence of preferential sampling is evaluated in situations where the researcher's goal is to optimize some objective function, e.g. reduce predictive variances. It will be shown through simulations that the optimal decision about this choice is substantially modified under a preferential sampling effect.

This paper is organized as follows: Section 2 presents some background about spatial processes and a methodology for fully Bayesian inference, and spatial prediction in Geostatistics under preferential sampling using MCMC methods is described. Section 3 presents a method to obtain the optimal design based on maximization of utility functions. Section 4 combines this methodology to obtain the optimal design under preferential sampling. In Sections 5 and 6, we use the methodology to obtain the optimal design in some simulated examples, and we analyse a real dataset where the researcher is interested in monitoring the occurrence of extreme events. Finally, Section 7 discusses the results obtained.

2 Spatial Process & Optimal Design

This section presents some background on Geostatistics and point processes, and presents a procedure for obtaining the optimal design through utility functions.

2.1 Geostatistics

Geostatistics deals with stochastic processes defined in a region D , $D \subset \mathbb{R}^k$, where usually $k = 1$ or 2 . Following the approach in Diggle et al. (1998), one can assume that the researcher is interested in studying the features of the stochastic process $\{S(x) : x \in D\}$. Additionally, assume that $S(x)$ is a Gaussian stationary and isotropic process, with zero mean, constant variance σ^2 and autocorrelation function $\rho(S(x), S(x+h); \phi) = \rho(\|h\|; \phi)$, $\forall x \in D$, which may depend on one or more parameters represented by ϕ . Several families of autocorrelation function can be found in the literature (see Cressie, 1993; Diggle and Ribeiro, 2007).

Assuming that n observations $Y_i = Y(x_i)$, $i = 1, \dots, n$ are available and

$$\begin{aligned} Y_i &= \mu + S(x_i) + Z_i, \\ E[Z_i] &= 0, \text{Var}[Z_i] = \tau^2, \quad \forall i, \end{aligned}$$

one can consider $\mathbf{Y} = (Y_1, \dots, Y_n)$ as a noisy version of the underlying process S . As usual, it will be assumed that $\mathbf{Z} = (Z_1, \dots, Z_n)$ has a Gaussian distribution independent of the process S .

Under the Bayesian paradigm, the posterior distribution of all model parameters must be obtained to make inference. Assuming that $\theta = (\tau^2, \sigma^2, \phi, \mu)$ represents the set of unknown parameters and that $p(\mathbf{y} | \theta)$ is the likelihood function, we need to specify a prior distribution $p(\theta)$ to obtain the posterior distribution $p(\theta | \mathbf{y})$. It is usual to assign Gamma distributions to the parameters ϕ , σ^{-2} and τ^{-2} and a Gaussian distribution to μ . Since the posterior distributions obtained have no closed form, we can approximate them using MCMC methods.

Usually there is an additional interest in obtaining predictions of the process S at locations not observed. Computations required in inference are usually implemented via discretization of the region D in M subregions or cells. Thus, we can redefine the underlying Gaussian process over the centroids of these subregions, $S = \{S_1, \dots, S_M\}$. In addition, we define the partition $S = \{S_{\mathbf{y}}, S_N\}$ to distinguish the underlying process associated with the n observed and the $N = M - n$ unobserved locations. The distribution of S is a multivariate Gaussian distribution with dimension $M = n + N$, mean vector $\mathbf{0}$ and autocorrelation matrix given by

$$R = \begin{pmatrix} R_n & R_{n,N} \\ R_{N,n} & R_N \end{pmatrix},$$

whose elements are defined by the autocorrelation function $\rho(\cdot; \phi)$. Then, we have the model

$$[\mathbf{y} | S_{\mathbf{y}}, \mu, \tau^{-2}] \sim N(\mathbf{1}\mu + S_{\mathbf{y}}, \tau^2 I_n), \quad (1)$$

$$[S | \sigma^{-2}, \phi] \sim N(\mathbf{0}, \sigma^2 R) \quad (2)$$

that is completed with the priors $\mu \sim N(0, k)$, $\tau^{-2} \sim G(a_\tau, b_\tau)$ and $\sigma^{-2} \sim G(a_\sigma, b_\sigma)$, where $k, a_\tau, b_\tau, a_\sigma$ and b_σ are known hyperparameters.

Without loss of generality with respect to the objectives of this work, the exponential autocorrelation function $\rho(\|h\|; \phi) = \exp(-\|h\|/\phi)$ will be used in the sequel.

The full conditional distributions for the parameters μ, τ^{-2} , σ^{-2} , and S can be updated by Gibbs sampling steps, whereas the range parameter ϕ can be updated by Metropolis steps in the MCMC algorithm. The full conditional distribution for S is of particular interest in Geostatistics and has Gaussian distribution with mean vector and covariance matrix given by

$$\sigma^2 R_{N,n} (\tau^2 I_n + \sigma^2 R_n)^{-1} (\mathbf{y} - \mathbf{1}\mu)$$

and

$$\sigma^2 R_N - \sigma^2 R_{N,n} (\tau^2 I_n + \sigma^2 R_n)^{-1} \sigma^2 R_{n,N}.$$

Expressions above are known as *kriging predictor* and *kriging variance*, respectively. There are extensions of the basic model of Geostatistics, most of them developed to deal with non-stationarity (Higdon et al. (1999); Fuentes and Smith (2001); Fuentes (2002) and Bornn et al. (2012)) or non-Gaussianity (Diggle et al., 1998).

2.2 Spatial Point Process

The use of point processes for modelling patterns of points in space has been intensified in the last decades, specially after the publication of the classic texts of Ripley (2005) and Diggle (1983). The study of point processes has also evolved, based in recent computational advances. Møller and Waagepetersen (2007) present an excellent review of the methods and models to spatial point processes and highlight several applications and computational aspects related to inference.

A spatial point process \mathbf{X} can be understood as a random finite subset of event locations belonging to a certain limited region $D \subset \mathbb{R}^k$, where usually $k = 1, 2$ or 3 . A spatial point process, defined in a region $D \subset \mathbb{R}^k$, governed by a non-negative random function $\Lambda = \{\Lambda(x) : x \in \mathbb{R}^k\}$, is a *Cox process* if the conditional distribution of $[\mathbf{X} \mid \Lambda = \lambda]$ is a Poisson process with intensity function $\lambda(x)$. Additionally, if one can assume that $\Lambda(x) = \exp\{Z(x)\}$, where $Z = \{Z(x) \in \mathbb{R}^k\}$ is a stationary and isotropic Gaussian process, then it is said to be *log-Gaussian Cox process* (Møller et al., 1998).

The likelihood function associated with this process is given by

$$p(\mathbf{x} \mid Z) \propto \exp\left(-\int_D \exp\{Z(x)\}dx\right) \prod_{i=1}^n \exp\{Z(x_i)\}.$$

Although this function is not analytically tractable, inference on a log-Gaussian Cox process can be performed in a reasonably simple way through Monte Carlo simulation methods (Møller et al., 1998). Again, it is usual to represent the domain D of the point process as a grid and approximate the Gaussian process $Z(x)$ by a finite-dimensional normal distribution defined on the grid. Waagepetersen (2004) showed that the discretized posterior for log-Gaussian Cox process converges to the exact posterior when the sizes of the grid cells tend to zero. According to Møller et al. (1998), if the point process is reasonably aggregated and has moderate intensity, the choice of the grid does not need to be fine to produce good results in inference. The inference conducted under the Bayesian paradigm, using MCMC methods, makes it relatively easy to obtain approximations of $p(\mathbf{x} \mid Z)$. On the other hand, the computational cost may be high (Møller and Waagepetersen, 2007).

2.3 Geostatistics under Preferential Sampling

In the Geostatistics literature, it is common to consider the sample points \mathbf{x} as fixed or, if coming from a stochastic process, independent of the process $S(x)$. When the sample design is stochastic, we must specify the joint distribution of $[Y, S, X]$. We have a process under *preferential sampling* if $[S, X] \neq [S][X]$, i.e. the sampling design is dependent of the spatial process.

The class of models proposed by Diggle et al. (2010) to accommodate preferential sampling effect assumes that, conditional on S , X is a *log-Gaussian Cox process* with intensity $\lambda(x) = \exp\{\alpha + \beta S(x)\}$. In addition, conditional on $[S, X]$, we have that $Y_i \sim N[\mu + S(x_i), \tau^2]$, $i = 1, \dots, n$. Diggle et al. (2010) presents a way to evaluate

this distribution through a fine discretization of the region D . The region D can be discretized into M cells with centroids $x_i, i = 1, \dots, M$, where only one point is expected in each cell.

In a Bayesian approach, we need to obtain the posterior distributions to make inference about the model parameters. For this purpose we assign Gaussian priors to α and β . Pati et al. (2011) proved that the use of improper priors for the parameter that controls the preferential sampling effects produces proper posteriors. Thus, the data provide enough information to perform inference for this model parameter, even under vague prior information.

Since the posterior distributions of these parameters have no closed form, we use an approximation making use of MCMC methods in a discretized version of the model given by

$$p(\mathbf{x} \mid S, \alpha, \beta) \propto \prod_{i=1}^M [\exp(\alpha + \beta S(x_i))]^{n_i} \exp\left(-\sum_i \Delta_i \exp(\alpha + \beta S(x_i))\right),$$

where n_i and Δ_i represent the counts and the volume of the i th subregion, $i = 1, \dots, M$.

The full conditional distributions of the parameters μ, ϕ, τ^{-2} and σ^{-2} are given by the same expressions obtained in the case of non-preferential sampling. The full conditional distributions of β and α are updated in a similar way. The full conditional distribution of S is updated in MCMC by Metropolis steps (see details in Appendix A).

3 Optimal Design

In general, finding an optimal design involves procedures for obtaining the maximum or minimum of an objective function. These objective functions usually quantify the gains and losses related to each possible decision. In this case, we need to decide which locations (in time or space) will be collected in order to better understand certain characteristics of the phenomenon. When the phenomenon of interest is studied assuming that it is governed by an underlying stochastic process, the methodology for obtaining the optimal design is usually performed via Decision Theory. For more details related to Decision Theory, see the classical textbook of DeGroot (2005).

According to Müller (1999), the procedure for obtaining the optimal design can be performed by defining a utility function $u(\mathbf{d}, \theta, \mathbf{y}_d)$, where $\mathbf{d} = (d_1, \dots, d_m)'$ represents the m new sample locations, $d_i \in D$, and $\mathbf{y}_d = (y_{d_1}, \dots, y_{d_m})'$ is the vector of future observations arising from it, $i = 1, \dots, m$. After a set of observations \mathbf{y} is available, the optimal design is the vector \mathbf{d}^* that maximizes the function

$$U(\mathbf{d}) = \int u(\mathbf{d}, \theta, \mathbf{y}_d) p_d(\mathbf{y}_d \mid \theta, \mathbf{y}) p(\theta \mid \mathbf{y}) d\theta d\mathbf{y}_d = E_{\theta, \mathbf{y}_d \mid \mathbf{y}}[u(\mathbf{d}, \theta, \mathbf{y}_d)].$$

In other approaches, one can include additional information to obtain the optimal design. In order to achieve the optimal design in time, Stroud et al. (2001) used covariates

to reduce the model parameter uncertainties, thereby obtaining the more appropriate point in time for the return of a patient undergoing treatment. On the other hand, Ding et al. (2008) used a hierarchical model to relate the effects of different treatments in clinical studies to determine the optimal design associated with a specific treatment.

An interesting strategy to optimize the expected utility is known as *an augmented model* (Müller, 1999). In this case, the optimal design point is considered as a parameter and can be estimated through its posterior distribution. Thus, an artificial probability model $h(\mathbf{d}, \theta, \mathbf{y}_d)$ is defined assuming that D is bounded and $u(\mathbf{d}, \theta, \mathbf{y}_d)$ is non-negative and limited. The distribution $h(\mathbf{d}, \theta, \mathbf{y}_d)$ is given by

$$h(\mathbf{d}, \theta, \mathbf{y}_d) \propto u(\mathbf{d}, \theta, \mathbf{y}_d) p_d(\mathbf{y}_d | \theta) p(\theta).$$

Under this assumptions, the marginal distribution of \mathbf{d} is proportional to

$$h(\mathbf{d}) \propto \int u(\mathbf{d}, \theta, \mathbf{y}_d) p_d(\mathbf{y}_d | \theta) p(\theta) d\theta d\mathbf{y}_d = U(\mathbf{d}).$$

Therefore, finding the mode of $h(\mathbf{d})$ becomes equivalent to maximizing the expected utility of \mathbf{d} . Since the sampled values of \mathbf{d} in the MCMC algorithm will be concentrated near regions of high expected utility, less time is consumed in the simulation procedure. In other words, one only needs to find the mode of the distribution, instead of dealing with an optimization problem.

It is important to remark that this methodology can be naturally performed in situations where one wants to choose $m > 1$ new design points. The main difficulty in this case is to face the increased computational cost involved since the evaluation of $u(\mathbf{d})$ is required at each iteration in MCMC. In addition, a large sample of $\mathbf{d} = (d_1, \dots, d_m)'$ would be needed to evaluate $h(\mathbf{d})$ properly. Alternatively, it is also possible to obtain the m new points sequentially. However, in this case, there is no guarantee that the resulting design will be optimal. The evaluation of the utility function also allows removing points instead of adding new points. In this case, the decision would be based on gains and losses in the utility after eliminating each site or group of sites.

3.1 Utility Functions

Commonly in Geostatistics, the researcher focuses more on spatial prediction than on inference about its parameters. In these cases, the choice of a utility function that depends on the predicted values and their variances is a natural choice. In this sense, the prediction variance associated with the location $x \in D$, $V(S(x) | \mathbf{y})$ gives to the researcher the degree of uncertainty about his/her predictions after observing the data \mathbf{y} . Thus, reducing this variance throughout the region D constitutes a reasonable criterion for choosing a new sample location d , $d \in D$. Based on this principle, the following utility function is a natural choice for implementing the approach described in Müller (1999) in the context of Geostatistics:

$$u(\mathbf{d}, \theta, \mathbf{y}_d) = \int [V(S(x) | \theta, \mathbf{y}) - V(S(x) | \theta, \mathbf{y}, \mathbf{y}_d)] dx, \quad (3)$$

which can be interpreted as the gain obtained in reduction of prediction uncertainty after observing \mathbf{y}_d . Thus we need to maximize its expectation given by $U(\mathbf{d})$.

Finally, it is also important to note that the predictive variances used in the utility function $u(\mathbf{d}, \theta, \mathbf{y}_d)$ depend only on the location of \mathbf{y}_d in D , instead of its value. This feature of the predictive variance avoids the necessity of evaluating all possible values of \mathbf{y}_d while maximizing $U(\mathbf{d})$, thus reducing the computational cost involved. More details about the evaluation of the utility function (3) are presented in Appendix B.

Other utility functions could be chosen. In Section 6, we employed a utility function that is higher for locations where extreme values or exceedances are expected. In this case, we have

$$u(d, \theta, y_d) = P[|S(x_d)| > x_0 \mid \theta, \mathbf{y}] \quad (4)$$

where x_d represents the location associated with y_d and x_0 is an extreme threshold. In practical situations, this expression could be improved by considering costs and risks related to the occurrences.

4 Optimal Design under Preferential Sampling

The problem of obtaining the optimal design \mathbf{d}^* for spatial processes under preferential sampling can be performed based on optimization of

$$U(\mathbf{d}) = E_{\theta, \mathbf{y}_d \mid \mathbf{x}, \mathbf{y}}[u(\mathbf{d}, \theta, \mathbf{y}_d)] = \int u(\mathbf{d}, \theta, \mathbf{y}_d) p(\mathbf{y}_d \mid \theta, \mathbf{x}, \mathbf{y}) p(\theta \mid \mathbf{x}, \mathbf{y}) d\theta d\mathbf{y}_d$$

where $p(\theta \mid \mathbf{x}, \mathbf{y})$ is obtained in Section 2. Given the posterior distribution of θ and the utility function, we can achieve the optimal design by seeking the mode of the posterior pseudo-distribution of \mathbf{d} . The greatest difficulty in this step is evaluate the effects of preferential sampling in the utility function $u(\mathbf{d}, \theta, \mathbf{y}_d)$. In many cases, like the situations presented in this paper, the evaluation of this function becomes challenging.

As will be noted in the next section, preferential sampling directly impacts on the estimation of the mean μ of the underlying Gaussian process. If the utility function $u(\mathbf{d}, \theta, \mathbf{y}_d)$ depends on this parameter, the choice of optimal design will be greatly affected. On the other hand, it would be expected that preferential sampling will also affects utility functions defined to quantify reductions of uncertainty related to each choice. This actually happens since the spatial configuration of the sample points also provides information about the underlying process.

Using as an example the utility function defined in Section 3, one would need to include the information provided by the observed point process \mathbf{x} , i.e.

$$u(\mathbf{d}, \theta, \mathbf{y}_d) = \int [V(S(x) \mid \theta, \mathbf{y}, \mathbf{x}) - V(S(x) \mid \theta, \mathbf{y}, \mathbf{x}, \mathbf{y}_d)] dx,$$

and one would still need to know the variance of the distribution of $[S \mid \theta, \mathbf{y}, \mathbf{x}]$. Samples of this distribution are easily obtained during the implementation of MCMC, as described in Section 2, but we cannot directly obtain estimates of this variance at each

iteration of the algorithm. To deal with this difficulty, one must resort to an approximation. A sampling-based approximation would require an additional MCMC sub-chain at each iteration, thus increasing substantially the already high computational cost. An analytic, cheaper alternative is to use a Gaussian approximation of this distribution in order to evaluate the variances required in the utility function (3). Note that this approximation is only used to evaluate $u(\mathbf{d}, \theta, \mathbf{y}_d)$. More detail about this approximation are presented in Appendix C. On the other hand, if the utility function (4) is used, we can directly evaluate $U(d)$ from $[S \mid \mathbf{y}, \mathbf{x}]$, since a sample of this distribution will be available after performing the MCMC.

5 Simulation Study

In this section, we simulate datasets according the model presented in Section 2.3 to illustrate the effects of preferential sampling on inference and optimal design choice. In all cases, we chose a set of parameters which do not produce very large samples, in order to enhance the effects of preferential sampling. We also consider situations where one needs to add $m = 1$ and $m = 2$ new locations to the sample design. In order to produce the observations y_1, \dots, y_n , we first generate a surface $S(x)$, according the Geostatistical model presented in Section 2.1, over the discretized region D . Conditioning on S , we then generate a point process from the log-Cox Gaussian model presented in Section 2.2 in order to obtain n observations. We considered five simulated studies:

- Case I. Simulation considering a one-dimensional region $D = [0, 100]$, partitioned into $M = 100$ sub-regions, where $(\alpha; \beta; \mu; \sigma^2; \phi; \tau^2; n) = (-3; 2; 12; 2; 20; 0.1; 18)$.
- Case II. Simulation considering a one-dimensional region $D = [0, 100]$, partitioned into $M = 200$ sub-regions, where $(\alpha; \beta; \mu; \sigma^2; \phi; \tau^2; n) = (-3.5; 3; 12; 1; 20; 0.01; 123)$.
- Case III. Simulation considering a one-dimensional region $D = [0, 200]$, partitioned into $M = 200$ sub-regions, where $(\alpha; \beta; \mu; \sigma^2; \phi; \tau^2; n) = (-1.5; 0.5; 12; 1; 20; 0.01; 56)$.
- Case IV. Simulation considering a two-dimensional region $D = [0, 100]$, partitioned into $M = 225$ sub-regions, where $(\alpha; \beta; \mu; \sigma^2; \phi; \tau^2; n) = (-8; 2; 12; 2; 20; 0.1; 12)$.
- Case V. Exactly as Case IV but with $M = 400$ sub-regions.

Case V was considered in order to evaluate the discretization effect. Figure 1 shows the Gaussian processes S simulated for Cases I–IV. Since $\beta > 0$ in all simulations, the observations are concentrated near the sites where the process S showed higher values.

5.1 Inference and Prediction

Posterior inference about model parameters was performed. Inference was also performed without considering preferential sampling, i.e. using the Geostatistical model of Section 2.1 to allow comparison. Prior distributions $\mu \sim \alpha \sim \beta \sim N(0; 10^3)$,

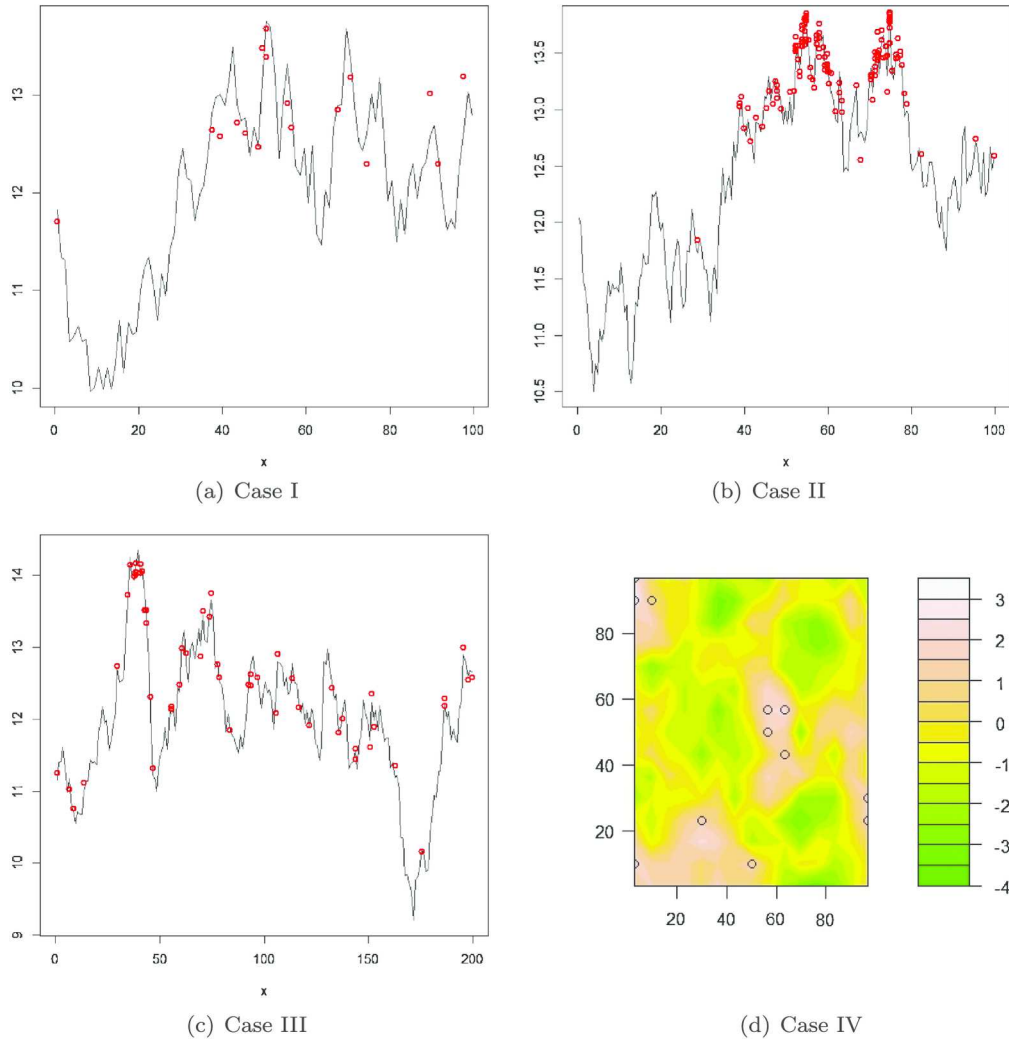


Figure 1: Realization of the one-dimensional (Cases I, II and III) and the two-dimensional (Case IV) simulated processes with respective observed values (circles). Except for (c), the observations are concentrated near the sites where the process S shows higher values.

$\tau^{-2} \sim \sigma^{-2} \sim G(2; 0.5)$ and $\phi \sim G(2; 0.05)$ were used in all cases. Furthermore, in *Cases I–III*, 500,000 iterations were generated in the MCMC algorithm and only the last 100,000 were used to compose the posterior distribution samples of the model parameters. For the other cases, 400,000 iterations were generated and only the last 50,000 were considered. The convergence of the chains was assessed by visual inspection of several chains generated from different initial values.

Considering the effect of preferential sampling, the posterior distributions of each model parameters are concentrated around the true values of the model's parameters, except for the σ^2 in *Case I*. However, this parameter was also underestimated considering a non-preferential model. Figure 2 shows that the variogram estimated assuming the preferential sampling effects is slightly closer to the true variogram for the *Cases I* and *V*. Figure 2 also shows that the variograms estimated by the preferential model are closer than those estimated by the non-preferential model in *Cases II* and *IV* (Figure 2(a),(b),(e), and (f)). Finally, in *Case III*, since β is small, the estimated variograms are very similar (also in Figure 2).

In particular, in *Case IV*, the posterior distributions are not concentrated around the true values for all parameters of the model. This difficulty is partly justified by the small sample size. However, the results obtained by the model with preferential sampling are more satisfactory. Besides a better estimated variogram, this also can be observed through the posterior distributions of μ , shown in Figure 3.

Figure 4 shows the predicted values of S , represented by the median of a posterior S , and the respective 95% credibility intervals for each model in *Cases I* and *II*.

Analysing Figure 4(a)–(b), we can see that only the credible intervals which consider the effect of preferential sampling encompass the most extreme values of the simulated process S . Additionally, it can be seen that the point estimates of S in the regions where the process was hardly observed are better with the model with preferential sampling. The differences are even more pronounced when we analyse the results for *Case II*, where the intervals are narrower for the preferential model (Figure 4(c)–(d)). In addition, the non-preferential model underestimates the process S .

Since the estimated variograms obtained by both models were similar in *Cases I* and *V*, it seems reasonable to conclude that the differences observed in prediction for these cases were caused mostly due to the differences between the predictive distributions $[S | \mathbf{y}]$ and $[S | \mathbf{y}, \mathbf{x}]$.

Finally, Figure 5 shows the predicted surfaces of $S + \mu$, represented by the posterior means, obtained for each model in *Case IV*. It can be observed that only the preferential model can identify regions where the underlying process presents low values. This feature avoid predictions concentrated around the mean, like those obtained by the traditional kriging methods. In addition, we can conclude that the small sample size intensifies the effects of preferential sampling in prediction.

All simulations presented in this section reflect situations where there are few observed sample points. This is not uncommon in practice, especially in monitoring studies of environmental or climatic phenomena which are rare or difficult to detect. However, even under these conditions, the use of models assuming the existence of preferential sampling effects has produced variogram and kriging surfaces generally closer to the true values when compared to the estimates produced without considering this effect. Another major advantage of using these models is the correction that is made during the inference about the underlying process mean. Finally, we showed the ability of

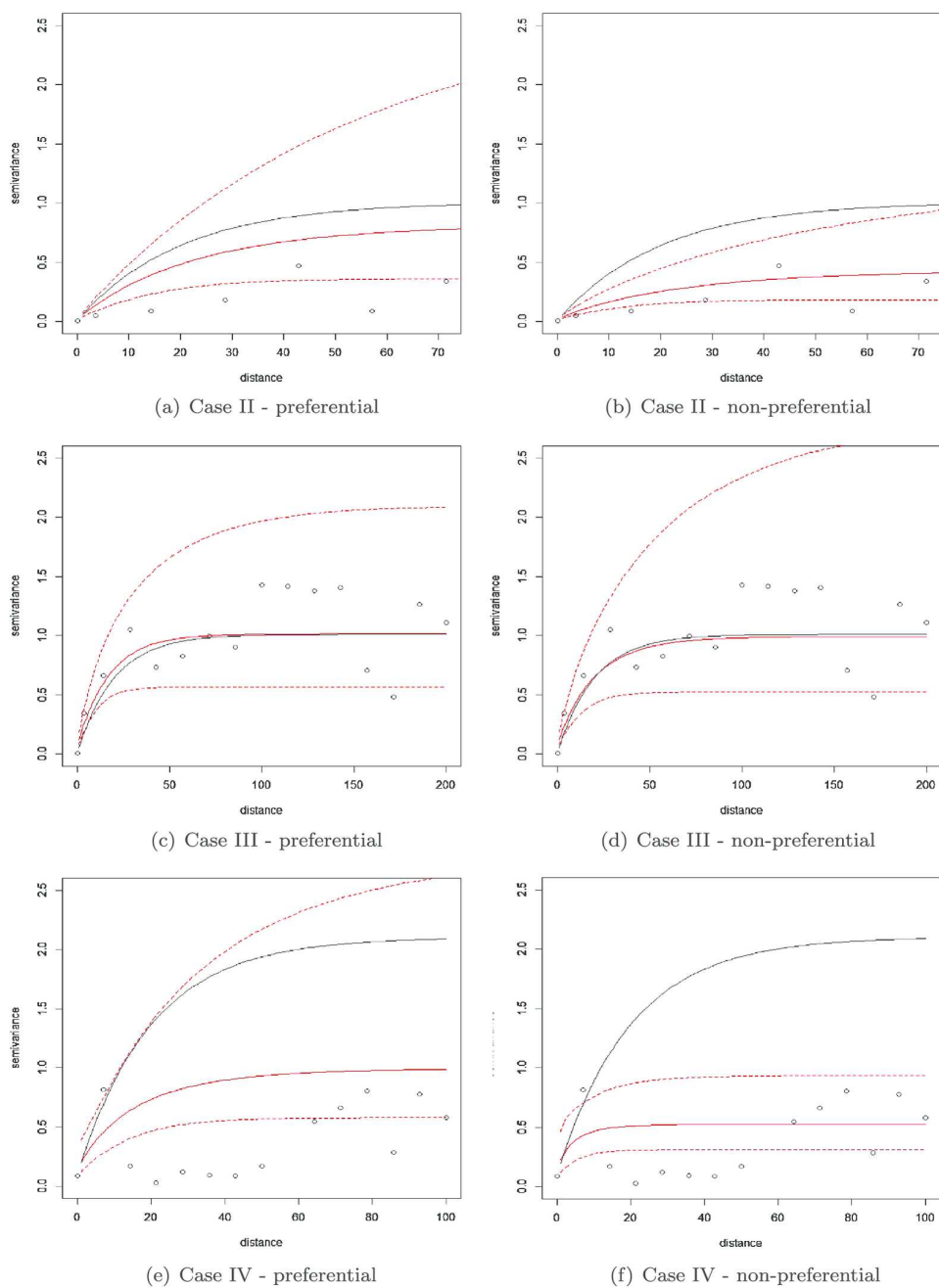


Figure 2: Posterior medians (red line) and respective 95% credibility intervals (dashed lines) of the variogram obtained by the preferential model (left) and by the non-preferential model (right) for Cases II, III and IV. The circles represent the empirical variogram and the black line represents the true variogram.

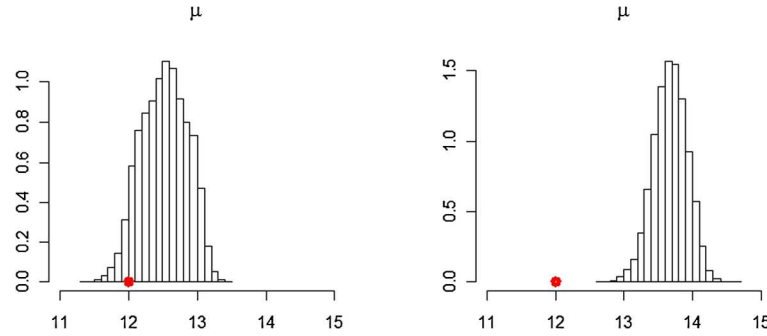


Figure 3: Posterior distributions of μ under preferential sampling (left) and without considering this effect (right) for *Case IV* (red points represent the true values of the parameters).

this model to identify areas where the underlying process takes extreme values, even in situations where there are no samples nearby.

Diggle et al. (2010) evaluated the influence of preferential sampling influence in the predictions of Gaussian processes using $[S|\mathbf{y}, \hat{\theta}]$ after correcting the bias caused by preferential sampling. However, these predictions did not take into account the information provided by \mathbf{x} . In the context of environmental monitoring, Shaddick and Zidek (2014) also presented a methodology for correcting the bias caused by a selective reduction of sample sites. In contrast, the Bayesian approach provides samples directly of the distribution of interest $[S|\mathbf{y}, \mathbf{x}]$. The comparison between the predictions assuming that θ is known reinforced the conclusion that methods based on corrections of the variogram's bias are not sufficient to reproduce the true uncertainty associated with the predictive distribution of the underlying process S .

Performing different simulated scenarios of prediction under preferential sampling effects, Gelfand et al. (2012) also concluded that they affect more significantly the spatial prediction than the estimation of the model parameters. In their paper, they discussed ways to evaluate the effects of preferential sampling by comparing two predicted surfaces. One of the forms of global comparison they mentioned is associated with the local and the global mean squared prediction error. The *Local Prediction Error* associated with x_0 , denoted $LPE(x_0)$, is given by

$$LPE(x_0) = E[\hat{S}(x_0) - S(x_0)]^2,$$

where $\hat{S}(x_0)$ is the predictor of S in x_0 . The *Global Prediction Error* is given by

$$GPE = \frac{1}{|D|} \int_D LPE(x) dx.$$

Table 1 presents the GPE values for each of the simulations. Based on this table, one can observe a significant reduction obtained by considering the effects of preferential sampling.

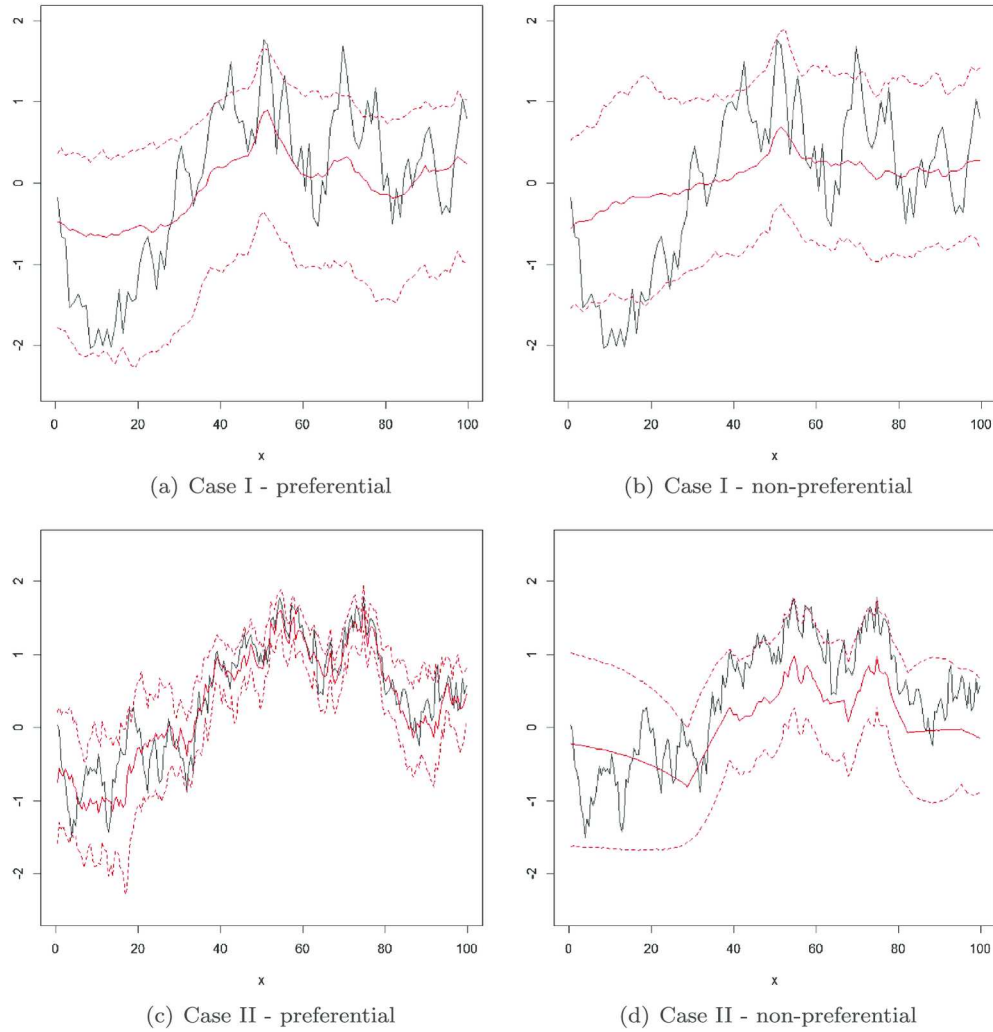


Figure 4: Simulated process (solid line), posterior median of the predictive distribution (red line) and the respective 95% credibility intervals (dashed lines) for S obtained by considering (left) and by not considering (right) the effect of preferential sampling in Cases I and II.

The one-dimensional simulation showed that the LPE s are reduced when they refer to the locations where the underlying process S has lower values. Similar conclusions were obtained in the second simulation, since errors remain smaller in regions where the magnitude of S is lower when the preferential sampling effect is taken into account in the modelling. Further simulations without the effect of preferential sampling were also performed. In such cases, in general, the model assuming preferential sampling produced posterior distributions of β centred at zero.

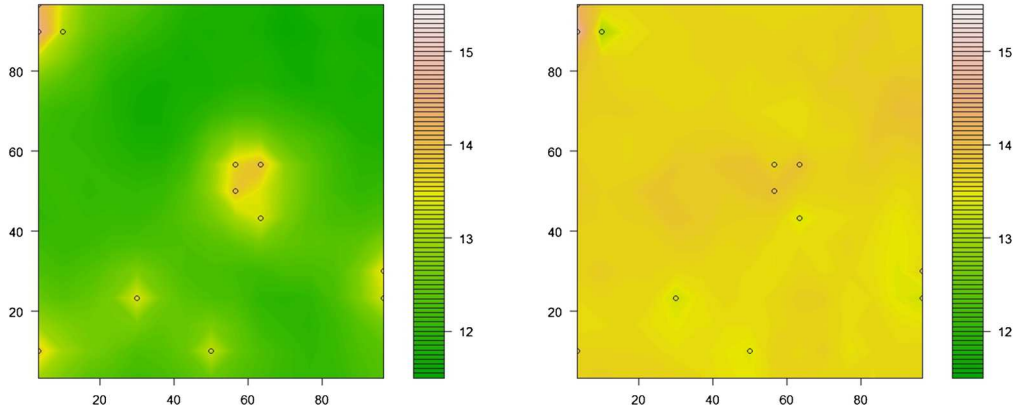


Figure 5: Posterior mean of the predictive distribution of $[S + \mu]$ obtained considering (left) and without considering (right) the effect of preferential sampling for *Case IV*.

Simulation	No preferential sampling	Preferential sampling
<i>Case I</i>	0.9496	0.7301
<i>Case II</i>	0.5533	0.1783
<i>Case III</i>	0.5286	0.4512
<i>Case IV</i>	2.1336	1.6789
<i>Case V</i>	1.6214	1.3482

Table 1: Global Prediction Error (GPE) for each of the simulations.

5.2 Optimal Design

Using the one-dimensional simulated data presented in *Case I*, 83 auxiliary points, required for evaluation of the utility function (3) described in Section 3, were also used to form a grid and obtaining the predictive variance reductions. Using the samples of θ , which were obtained from the posterior distribution in the MCMC algorithm, we can generate samples of d as mentioned in Section 3 to obtain a new optimal sample point for the case where $m = 1$.

Figure 6 shows the histograms of the posterior pseudo-distribution of d with and without the preferential sampling assumption, respectively. It can be noted that the optimal design choice without preferential sampling leads the researcher to select locations where there are no nearby points, i.e. in the interval $[5, 35]$. On the other hand, under preferential sampling, the results lead the researcher to a different direction. In this case, except in the subregions where there are several observed samples, the other choices have similar expected utilities. In summary, under preferential sampling, the optimal design choice is notably changed when the researcher wants to reduce the predictive variance.

To illustrate the effect of the preferential sampling where $m > 1$, we obtained the pseudo-distribution of $\mathbf{d} = (d_1, d_2)'$ assuming that two new locations would need to be added in the sample design. The results are presented in Figure 7. Analysing the results

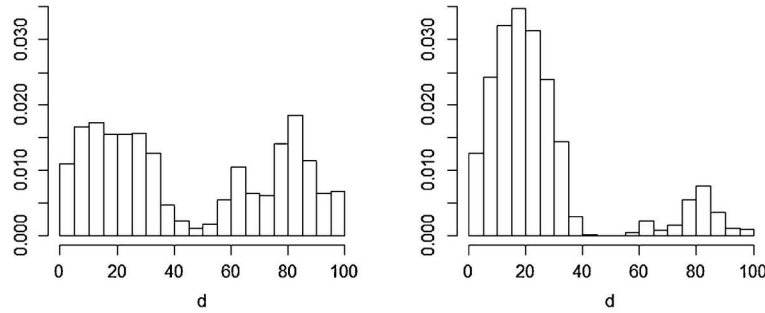


Figure 6: Posterior pseudo-distributions of d under the effect of preferential sampling (left) and without considering this effect (right) for *Case I*.

for the non-preferential model, it can be seen that the optimal solution could be to add one point d_1 from interval $[0, 20]$ and to make no restrictions to the location of the other point d_2 , since $h(\mathbf{d})$ does not vary much. On the other hand, the results for the preferential model indicate that the optimal design should include one point d_1 from interval $[0, 4]$ and another point d_2 from interval $[88, 92]$.

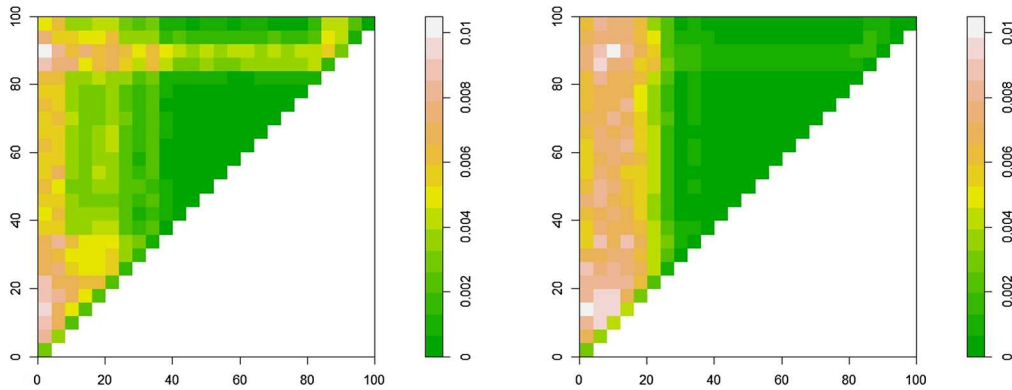


Figure 7: Posterior pseudo-distributions of $\mathbf{d} = (d_1, d_2)'$ under the effect of preferential sampling (left) and without considering this effect (right) for *Case I*.

In the two-dimensional *Case IV*, 900 auxiliary points were also used to form a grid and obtaining the predictive variance reductions, as in the previous section. Figure 8 shows the posterior pseudo-distributions of d under preferential sampling and without considering this effect for the case where $m = 1$.

It can be noted in Figure 8 that the areas with the highest expected utility are not the same for the two models. As expected, the results obtained without considering the effect of preferential sampling makes the researcher choose locations far from the observed points. In contrast, under preferential sampling, the largest utilities expected are more dispersed over region D . Again, under preferential sampling, the optimal design choice

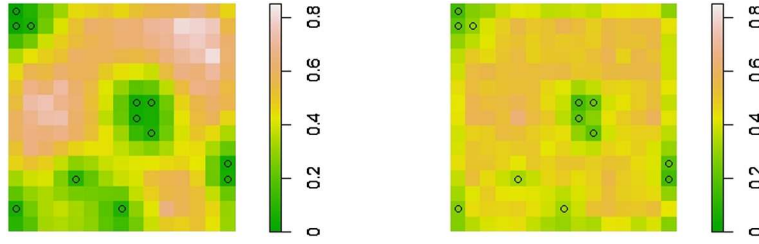


Figure 8: Posterior pseudo-distribution of d under the effect of preferential sampling (left) and without considering this effect (right) for *Case IV*. The densities of these pseudo-distribution are multiplied by 100 for better visualization.

was changed, since the regions with few sample sites also provide useful information about S .

The two simulated studies have involved situations where the inference step has produced similar results (in the one-dimensional case) and different results (in the two-dimensional case). Surprisingly, even in the one-dimensional case, where the estimated variograms were similar, the process of choosing the optimal design led to quite different results.

However, the use of a different utility function can produce even more extreme results. Even though the utility function chosen in this paper was designed for our Geostatistics goals, other functions could be considered. According to the results obtained from the simulated studies, utility functions that depend on the underlying process μ may also be affected by the preferential sampling effect. This situation will be explored in next section. There are other effects that may affect the results, such as the choice of the auxiliary grid (used to evaluate the predictive variance reduction) and the discretization level of D . Current computational costs associated with this methodology are still a barrier for a more thorough evaluation of the degree of influence of each of these marginal effects.

Effectiveness of the Optimal Decision

After obtaining the optimal design, one can evaluate if the results are better under preferential sampling. Thus, we performed an analysis of the GPE after this decision to the two-dimensional simulated data *Case IV*. The coordinates of the optimal design was $x_d = (90.00; 76.66)$, under preferential sampling, and $x_d = (30.00; 50.00)$ without this effect. In addition, it was assumed that $\tau^2 = 0$.

Finally, we proceed to the inference via classical Geostatistics methods but including the optimal data point Y_d obtained under preferential sampling. The results provided a $GPE = 1.8392$, which is lower than that obtained using the optimal data point Y_d pointed out by the non-preferential model ($GPE = 2.3348$). Thus, the optimal design obtaining under preferential sampling was more advantageous even when the inference is performed via classical Geostatistics methods.

6 Case Study: Rainfall Data in Rio de Janeiro

The methodology is now applied to a real scenario in the context of monitoring networks. More specifically, we will analyse pluviometric precipitation data obtained from 32 monitoring stations located in the city of Rio de Janeiro, Brazil. The data refer to the period from 1 to 31 October 2005 and were obtained from the *Pereira Passos* Institute, an official agency associated with the local government. The rainfall during this month, which begins the rainy season in the Brazil's Southeast, is of particular interest to meteorologists and government agencies (Alves et al., 2005). Figure 9 shows the map of the Rio de Janeiro city with the respective precipitation levels observed.

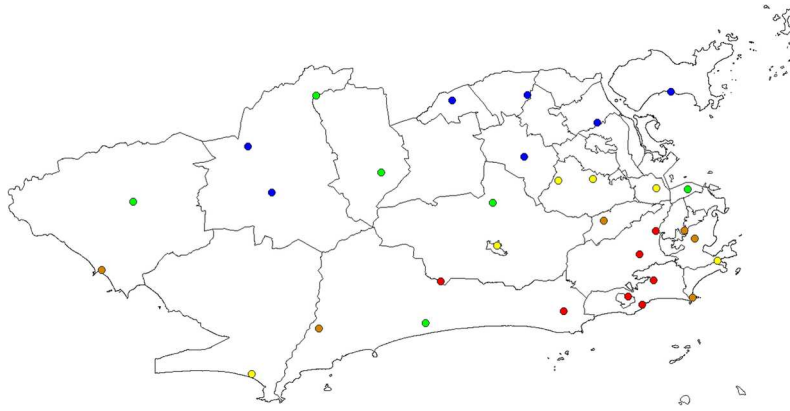


Figure 9: Pluviometric precipitation in Rio de Janeiro city in October/2005 (separated according the 0.20, 0.40, 0.60, and 0.80 quantiles and grouped by the colours: blue, green, yellow, brown, and red, respectively).

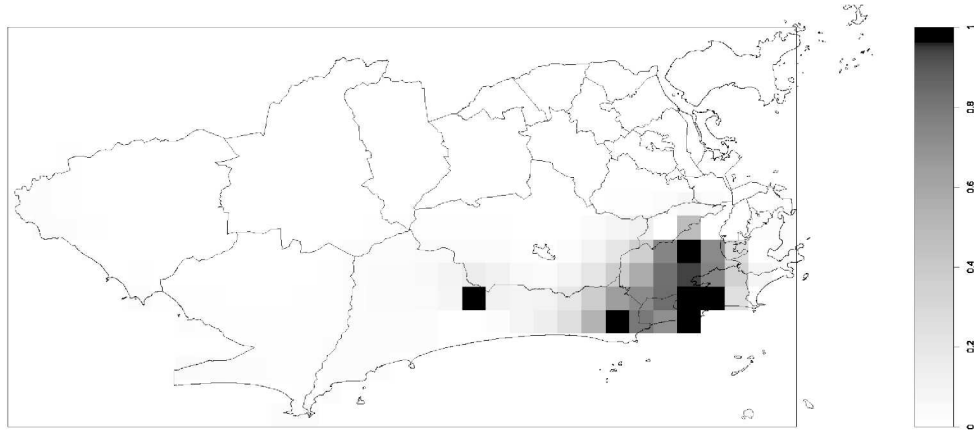
Analysis of the spatial distribution of the precipitation seems to indicate that the stations are more concentrated near places where rainfall level is higher. Even though the geography and the spatial distribution of economic activities could be considered as possible causes of this sample design, the methodology for choosing a new design point under the effect of preferential sampling can be employed here.

For inference, we used the same priors of the previous simulations (by changing some hyperparameters) and the study area was partitioned into $M = 332$ subregions. We monitored 100,000 iterations in the MCMC algorithm for both models and the first 10,000 were considered as burn-in. The convergence of the chains was assessed by visual inspection of several chains generated from different initial values. Table 2 presents summaries of the posterior distributions for all model parameters. An analysis of the results suggests that the effect of preferential sampling is significant, indicating a positive association between the sample design and the rainfall intensity.

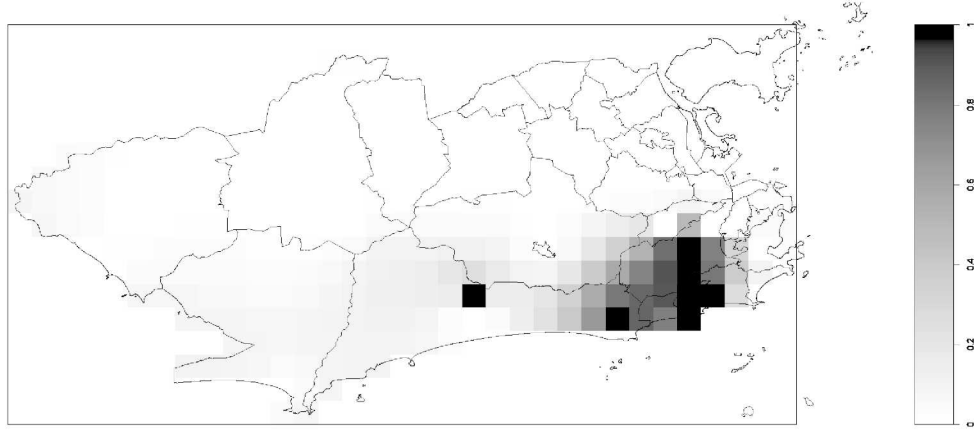
As for the other model parameters, we observed differences between the estimates for the mean μ , which can be explained by the presence of preferential sampling effects. The utility function (4) was used to obtaining the optimal design, with $x_0 = 200$ mm.

Model parameters	Preferential	Non-Preferential
τ^2	1.25 (0.49; 2.90)	1.32 (0.51; 3.45)
σ^2	4289.92 (2096.31; 10528.91)	4132.72 (2120.30; 8514.38)
μ	104.84 (97.73; 110.60)	119.88 (111.49; 130.32)
ϕ	10.69 (4.27; 26.75)	10.43 (4.51; 22.76)
α	-3.84 (-4.24; -3.48)	—
β	0.008 (0.002; 0.014)	—

Table 2: Posterior mean and 95% credibility interval (in parentheses) for model parameters.



(a) Preferential



(b) Non-Preferential

Figure 10: Expected utilities $U(d)$ obtained under preferential sampling (a) and without considering this effect (b) for the Rainfall data.

This utility function assigns greater utility for regions where it is most likely to observe a monthly rainfall above 200 mm. The expected utility $U(d)$ obtained for each model is shown in Figure 10.

The preferential model has concentrated high expected utility into a small region in map since the utility function favours regions with the highest probability of extremes values and due to the overestimation of μ . On the other hand, the non-preferential model has produced expected utilities more spread out over the southern part of the city.

7 Discussion

The results obtained in Sections 5 and 6 help us to understand the effects of preferential sampling on inference and optimal design choice in the context of Geostatistics. The results of Section 5 showed that the bias corrections made during the estimation step are not enough to ensure a satisfactory prediction. On the other hand, the Bayesian approach presented here allows us to make predictions about any functional of S directly from $[S \mid \mathbf{y}, \mathbf{x}]$, which is the correct predictive distribution under the effect of preferential sampling. Furthermore, the inference about the parameters that define the effect of preferential sampling, i.e. α and β , was quite satisfactory in all simulations. Despite the high associated computational cost, the approach is computationally feasible and showed satisfactory results.

In practical situations, it may be unlikely to assume that the design is governed by a log-Gaussian Cox process. However, this model seems to be flexible to obtain understanding about the consequences if the researcher has no covariates or a better explanation for its true causes. Although not applicable in the strictly spatial context, an alternative approach to detect and deal with preferential sampling is spatio-temporal analysis, since the changes on site locations over time may be informative about the association between its configuration and the underlying process. Zidek et al. (2014) present a method that can learn about the preferential selection process over time and that allows the researcher to deal with its effects.

Traditional approaches to deal with preferential sampling also include the fit of trend surfaces to assess first order effects and, in the context of survey sampling methods, weighting schemes. However, both require some prior knowledge about the sample selection processes. The researcher must carefully evaluate the (theoretical or empirical) evidence of preferential sampling effects. The use of this approach can produce misleading results otherwise. If there is evidence of preferential sampling, the authors believe the use of models based on latent processes, such as log-Gaussian Cox process, should be used even when the dependence structure between X and S is not completely known.

As evidenced by simulations, the effects of preferential sampling on optimal design choice cannot be disregarded. The knowledge about the sampling pattern reduces the predictive variance of S in areas poorly sampled, substantially changing the optimal decision. On the other hand, a utility function based on exceedances seems to overestimate μ under preferential sampling. As a result, the ideal region to receive a new

sample becomes very small. However, obtaining the optimal design in situations where there is a need to monitor extreme events or exceedances may be not simple. Chang et al. (2007) presented an approach to deal with some challenges arising in designing networks for monitoring fields of extremes, as the loss of spatial dependence and the limitations of conventional approaches.

Potential areas to apply this methodology include the study of phenomena scarcely observed and those in which, due to researcher's interest or limited resources, can only be observed in locations considered critical. One can include the monitoring of mosquitoes or other disease spreading pests that only become detectable in places where its occurrence is very high among the phenomena that are scarcely observed and difficult to be detected. Another potential application is the monitoring of maximum and minimal temperatures in regions near airports or industrial plants. In both cases, it seems reasonable to infer that the way the phenomenon is observed can be related to the underlying process. In this situation, the optimal design choice can change significantly, since the spatial point pattern brings valuable information to the researcher.

The methodology employed can be adjusted in order to produce pseudo-distributions of d more peaked around the mode. The strategy based on simulated annealing (Müller, 1999) can be explored for this purpose. The approach of Müller (1999) for optimal design choice has computational advantages in comparison with traditional procedures of optimization. The choice of a utility function based on predictive variance reduction is easily justified in Geostatistics. Under preferential sampling, the use of utility functions that directly depend on μ seems to be more affected than those based in variance reductions.

This methodology also has an expensive computational cost and alternative procedures can be used to deal with this problem, e.g. Simpson et al. (2011) that use the *Integrated Nested Laplace Approximation* – INLA methods (Rue et al., 2009), the use of Predictive Process (Banerjee et al., 2008), methods to approximate likelihoods (see Stein et al., 2004; Fuentes, 2007) and the use of sparse covariance matrices (Furrer et al., 2006).

The authors recognize the limitations in the simulation studies presented here. The complexity and the variety of situations arising from designing under preferential sampling lead us to focus on specific features rather than obtaining more general conclusions. Finally, the authors also invite researchers interested in reproducing the methodology presented in this paper to contact the authors in order to obtain more details about the computational implementation.

Appendix A

In the preferential model, the full conditional distributions of S , β and α are proportional to

$$p(S \mid \mu, \tau^{-2}, \sigma^{-2}, \phi, \alpha, \beta, \mathbf{x}, \mathbf{y}) \propto \exp \left\{ -\frac{1}{2\tau^2} [S'_y S_y - 2S'_y (\mathbf{y} - \mu \mathbf{1})] + \beta S'_n \mathbf{n} - \frac{S' R_M^{-1} S}{2\sigma^2} \right\} \\ \times \exp \left\{ -\Delta e^\alpha \sum_{i=1}^M \exp(\beta S(x_i)) \right\},$$

$$\begin{aligned}
p(\beta \mid S, \mu, \tau^{-2}, \sigma^{-2}, \phi, \alpha, \mathbf{x}, \mathbf{y}) &\propto p(\mathbf{x} \mid S, \alpha, \beta) p(\beta) \\
&\propto \exp \left\{ \beta S' \mathbf{n} - \Delta e^\alpha \sum_{i=1}^M \exp(\beta S(x_i)) - \frac{\beta^2}{2k} \right\}, \\
p(\alpha \mid S, \mu, \tau^{-2}, \sigma^{-2}, \phi, \beta, \mathbf{x}, \mathbf{y}) &\propto \exp \left\{ n\alpha - \Delta e^\alpha \sum_{i=1}^M \exp(\beta S(x_i)) - \frac{\alpha^2}{2k} \right\}.
\end{aligned}$$

In the MCMC, these quantities can be updated in Metropolis steps assuming a Gaussian proposal distribution centred in the previous values sampled. Then, we accept the proposal with probability

$$\begin{aligned}
p_S &= \exp \left\{ -\frac{1}{2\tau^2} [S_{\mathbf{y}}'^{prop} S_{\mathbf{y}}^{prop} - S_{\mathbf{y}}' S_{\mathbf{y}} - 2(S_{\mathbf{y}}'^{prop} - S_{\mathbf{y}})'(\mathbf{y} - \mu \mathbf{1})] + \beta(S^{prop} - S)' \mathbf{n} \right\} \\
&\quad \times \exp \left\{ \Delta e^\alpha \sum_{i=1}^M [\exp(\beta S(x_i)) - \exp(\beta S(x_i)^{prop})] + \frac{(S' R_M^{-1} S - S'^{prop} R_M^{-1} S^{prop})}{2\sigma^2} \right\}, \\
p_\beta &= \exp \left\{ (\beta^{prop} - \beta) S' \mathbf{n} + \Delta e^\alpha \sum_{i=1}^M (\exp(\beta S(x_i)) - \exp(\beta^{prop} S(x_i))) + \frac{(\beta^2 - \beta^{2prop})}{2k} \right\}, \\
p_\alpha &= \exp \left\{ (\alpha^{prop} - \alpha) n + (e^\alpha - e^{\alpha^{prop}}) \Delta \sum_{i=1}^M \exp(\beta S(x_i)) + \frac{(\alpha^2 - \alpha^{2prop})}{2k} \right\},
\end{aligned}$$

respectively. The vector $\mathbf{n}' = (n_1, n_2, \dots, n_M)$ represents the number of observations in each subregion, where $\sum_{i=1}^M n_i = n$.

In both models, the full conditional distribution of ϕ is proportional to

$$p(\phi \mid S, \mu, \tau^{-2}, \sigma^{-2}, \alpha, \beta, \mathbf{y}) \propto |R_M|^{-1/2} \phi^{a_\phi - 1} \exp \left\{ -\frac{S' R_M^{-1} S}{2\sigma^2} - b_\phi \phi \right\}$$

and this parameter can be updated in Metropolis steps assuming the following proposal distribution

$$q(\phi^{prop} \mid \phi) \sim \text{Lognormal}(\ln(\phi) - \delta/2; \delta),$$

where δ must be chosen in order to produce reasonable acceptance rates in MCMC. Then, we accept the proposal with probability

$$\begin{aligned}
p_\phi &= \left(\frac{|R_M|^{prop}}{|R_M|} \right)^{-1/2} \left(\frac{\phi^{prop}}{\phi} \right)^{a_\phi} \times \exp \left\{ -\frac{(S' R_M^{prop-1} S - S' R_M^{-1} S)}{2\sigma^2} + b_\phi(\phi - \phi^{prop}) \right. \\
&\quad \left. - \frac{(\ln \phi - \ln \phi^{prop} + \delta/2)^2 - (\ln \phi^{prop} - \ln \phi + \delta/2)^2}{2\delta} \right\}.
\end{aligned}$$

Appendix B

To evaluate the integral in expression of $u(\mathbf{d}, \theta, \mathbf{y}_d)$ a discretization of the region D can be applied yielding M subregions, as described in Section 2. Then, we have that

$u(\mathbf{d}, \theta, \mathbf{y}_d)$ can be approximated by

$$\tilde{u}(\mathbf{d}, \theta, \mathbf{y}_d) = \frac{1}{M} \sum_i [V(S_i | \theta, \mathbf{y}) - V(S_i | \theta, \mathbf{y}, \mathbf{y}_d)],$$

where $[S_i | \theta, \mathbf{y}]$ and $[S_i | \theta, \mathbf{y}, \mathbf{y}_d]$, $i = 1, \dots, M$, are Gaussian with respective means

$$\sigma^2 \mathbf{r}'_n (\tau^2 I_n + \sigma^2 R_n)^{-1} (\mathbf{y} - \mathbf{1}\mu) \quad \text{and} \quad \sigma^2 \mathbf{r}'_{n+m} (\tau^2 I_{n+m} + \sigma^2 R_{n+m})^{-1} (\mathbf{y}^* - \mathbf{1}\mu)$$

and variances

$$\sigma^2 - \sigma^2 \mathbf{r}'_n (\tau^2 I_n + \sigma^2 R_n)^{-1} \sigma^2 \mathbf{r}_n \quad \text{and} \quad \sigma^2 - \sigma^2 \mathbf{r}'_{n+m} (\tau^2 I_{n+m} + \sigma^2 R_{n+m})^{-1} \sigma^2 \mathbf{r}_{n+m},$$

where $\mathbf{y}^* = (\mathbf{y}, \mathbf{y}_d)'$.

Appendix C

Following the procedure proposed in Section 4, we can make an approximation in $p(\mathbf{x} | S, \alpha, \beta)$ to obtain an analytical expression of $V[S | \theta, \mathbf{y}, \mathbf{x}]$. In particular, expanding the exponential function in Taylor series around zero, up to the second order term, we obtain the following approximation

$$\sum_i \exp(\alpha + \beta S(x_i)) \approx e^\alpha \left(\mathbf{1}_M + \beta \mathbf{1}_M' S + \frac{\beta^2}{2} S' S \right).$$

By inserting this expression in $p(\mathbf{x} | S, \alpha, \beta)$, it can be shown that the conditional distribution $[S | \theta, \mathbf{y}, \mathbf{x}]$ becomes Gaussian with mean vector Θ and covariance matrix Σ given by

$$\Theta = \Sigma \times \begin{pmatrix} (\mathbf{y}_n - \mu \mathbf{n}) \tau^{-2} + \beta \mathbf{n} - \Delta \beta e^\alpha \mathbf{n} \\ -\Delta \beta e^\alpha \mathbf{1}_N \end{pmatrix} \quad \text{and}$$

$$\Sigma = \begin{pmatrix} (\tau^{-2} + \Delta \beta^2 e^\alpha) I_n + R_n^{-1} R_{n,N} A^{-1} R_{N,n} R_n^{-1} + \sigma^{-2} R_n^{-1} & -R_n^{-1} R_{n,N} A^{-1} \\ -A^{-1} R_{N,n} R_n^{-1} & \Delta \beta^2 e^\alpha I_N + A^{-1} \end{pmatrix}^{-1},$$

where $A = \sigma^2 R_N - \sigma^2 R_{N,n} R_n^{-1} R_{n,N}$, and the vectors \mathbf{n} and \mathbf{y}_n represent the number of observations and the total observed in each subregion of D , respectively. If $\beta = 0$, this matrix becomes equal to the kriging variance matrix traditionally obtained in Geostatistics.

References

- Alves, L., Marengo, J., Júnior, H., and Castro, C. (2005). “Beginning of the rainy season in southeastern Brazil: Part 1 – Observational studies (in Portuguese).” *Revista Brasileira de Meteorologia*, 20(3): 385–394. [727](#)

- Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). “Gaussian predictive process models for large spatial data sets.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4): 825–848. [MR2523906](#). doi: <http://dx.doi.org/10.1111/j.1467-9868.2008.00663.x>. 730
- Bornn, L., Shaddick, G., and Zidek, J. (2012). “Modeling nonstationary processes through dimension expansion.” *Journal of the American Statistical Association*, 107(497): 281–289. [MR2949359](#). doi: <http://dx.doi.org/10.1080/01621459.2011.646919>. 713
- Boukouvalas, A., Cornford, D., and Stehlik, M. (2009). “Approximately optimal experimental design for heteroscedastic Gaussian process models.” Technical report, Nonlinear Complexity Research Group (NCRG). 711
- Chang, H., Fu, A. Q., Le, N. D., and Zidek, J. V. (2007). “Designing environmental monitoring networks to measure extremes.” *Environmental and Ecological Statistics*, 14(3): 301–321. [MR2405332](#). doi: <http://dx.doi.org/10.1007/s10651-007-0020-5>. 730
- Cressie, N. (1993). *Statistics for spatial data*. John Wiley and Sons, Inc. [MR1239641](#). 711, 712
- DeGroot, M. H. (2005). *Optimal statistical decisions*, volume 82. Wiley-Interscience. [MR2288194](#). doi: <http://dx.doi.org/10.1002/0471729000>. 715
- Diggle, P. (1983). *Statistical Analysis of Spatial Point Patterns*. London: Academic Press. [MR0743593](#). 714
- Diggle, P. J. and Lophaven, S. (2006). “Bayesian geostatistical design.” *Scandinavian Journal of Statistics*, 33(1): 53–64. [MR2255109](#). doi: <http://dx.doi.org/10.1111/j.1467-9469.2005.00469.x>. 711
- Diggle, P. J., Menezes, R., and Su, T.-L. (2010). “Geostatistical inference under preferential sampling.” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(2): 191–232. [MR2744471](#). doi: <http://dx.doi.org/10.1111/j.1467-9876.2009.00701.x>. 711, 714, 722
- Diggle, P. J. and Ribeiro, P. J. (2007). *Model-based geostatistics*. Springer. [MR2293378](#). 712
- Diggle, P. J., Tawn, J., and Moyeed, R. (1998). “Model-based geostatistics.” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(3): 299–350. [MR1626544](#). doi: <http://dx.doi.org/10.1111/1467-9876.00113>. 711, 712, 713
- Ding, M., Rosner, G. L., and Müller, P. (2008). “Bayesian optimal design for phase II screening trials.” *Biometrics*, 64(3): 886–894. [MR2526640](#). doi: <http://dx.doi.org/10.1111/j.1541-0420.2007.00951.x>. 716
- Fernández, J., Real, C., Couto, J., Aboal, J., and Carballeira, A. (2005). “The effect of sampling design on extensive bryomonitoring surveys of air pollution.” *Science of the total environment*, 337(1): 11–21. 711

- Fuentes, M. (2002). “Spectral methods for nonstationary spatial processes.” *Biometrika*, 89(1): 197–210. MR1888368. doi: <http://dx.doi.org/10.1093/biomet/89.1.197>. 713
- (2007). “Approximate likelihood for large irregularly spaced spatial data.” *Journal of the American Statistical Association*, 102(477): 321–331. MR2345545. doi: <http://dx.doi.org/10.1198/016214506000000852>. 730
- Fuentes, M. and Smith, R. L. (2001). “A new class of nonstationary spatial models.” Technical report, North Carolina State University, Raleigh, NC. 713
- Furrer, R., Genton, M. G., and Nychka, D. (2006). “Covariance tapering for interpolation of large spatial datasets.” *Journal of Computational and Graphical Statistics*, 15(3). MR2291261. doi: <http://dx.doi.org/10.1198/106186006X132178>. 730
- Gelfand, A. E., Sahu, S. K., and Holland, D. M. (2012). “On the effect of preferential sampling in spatial prediction.” *Environmetrics*, 23(7): 565–578. MR3020075. doi: <http://dx.doi.org/10.1002/env.2169>. 722
- Gumprecht, D., Müller, W. G., and Rodríguez-Díaz, J. M. (2009). “Designs for detecting spatial dependence.” *Geographical Analysis*, 41(2): 127–143. 711
- Higdon, D., Swall, J., and Kern, J. (1999). “Non-stationary spatial modeling.” In: *Bayesian statistics* (J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith), 6(1): 761–768. 713
- Møller, J., Syversveen, A. R., and Waagepetersen, R. P. (1998). “Log-Gaussian Cox processes.” *Scandinavian Journal of Statistics*, 25(3): 451–482. MR1650019. doi: <http://dx.doi.org/10.1111/1467-9469.00115>. 711, 714
- Møller, J. and Waagepetersen, R. P. (2007). “Modern statistics for spatial point processes.” *Scandinavian Journal of Statistics*, 34(4): 643–684. MR2392447. doi: <http://dx.doi.org/10.1111/j.1467-9469.2007.00569.x>. 714
- Müller, P. (1999). “Simulation-based optimal design.” In: *Bayesian statistics* (J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith), 6: 459–474. MR1723509. 711, 712, 715, 716, 730
- Müller, P., Berry, D. A., Grieve, A. P., Smith, M., and Krams, M. (2007). “Simulation-based sequential Bayesian design.” *Journal of Statistical Planning and Inference*, 137(10): 3140–3150. MR2364157. doi: <http://dx.doi.org/10.1016/j.jspi.2006.05.021>. 711
- Müller, P., Sansó, B., and De Iorio, M. (2004). “Optimal Bayesian design by inhomogeneous Markov chain simulation.” *Journal of the American Statistical Association*, 99(467): 788–798. MR2090911. doi: <http://dx.doi.org/10.1198/016214504000001123>. 711
- Müller, W. G. and Stehlík, M. (2010). “Compound optimal spatial designs.” *Environmetrics*, 21(3-4): 354–364. MR2842248. doi: <http://dx.doi.org/10.1002/env.1009>. 711

- Pati, D., Reich, B. J., and Dunson, D. B. (2011). “Bayesian geostatistical modelling with informative sampling locations.” *Biometrika*, 98(1): 35–48. MR2804208. doi: <http://dx.doi.org/10.1093/biomet/asq067>. 715
- Ripley, B. D. (2005). *Spatial statistics*, volume 575. Wiley. MR0624436. 714
- Rue, H., Martino, S., and Chopin, N. (2009). “Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2): 319–392. MR2649602. doi: <http://dx.doi.org/10.1111/j.1467-9868.2008.00700.x>. 730
- Shaddick, G. and Zidek, J. V. (2014). “A case study in preferential sampling: Long term monitoring of air pollution in the UK.” *Spatial Statistics*, 9: 51–65. 722
- Simpson, D., Illian, J., Lindgren, F., Sørbye, S., and Rue, H. (2011). “Going off grid: Computationally efficient inference for log-Gaussian Cox processes.” arXiv:1111.0641. 730
- Stein, M. L., Chi, Z., and Welty, L. J. (2004). “Approximating likelihoods for large spatial data sets.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(2): 275–296. MR2062376. doi: <http://dx.doi.org/10.1046/j.1369-7412.2003.05512.x>. 730
- Stroud, J. R., Müller, P., and Rosner, G. L. (2001). “Optimal sampling times in population pharmacokinetic studies.” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(3): 345–359. MR1856330. doi: <http://dx.doi.org/10.1111/1467-9876.00239>. 715
- Waagepetersen, R. (2004). “Convergence of posteriors for discretized log Gaussian Cox processes.” *Statistics & Probability Letters*, 66(3): 229–235. MR2044908. doi: <http://dx.doi.org/10.1016/j.spl.2003.10.014>. 714
- Zhu, Z. and Stein, M. L. (2005). “Spatial sampling design for parameter estimation of the covariance function.” *Journal of Statistical Planning and Inference*, 134(2): 583–603. MR2200074. doi: <http://dx.doi.org/10.1016/j.jspi.2004.04.017>. 711
- Zidek, J. V., Shaddick, G., Taylor, C. G., et al. (2014). “Reducing estimation bias in adaptively changing monitoring networks with preferential site selection.” *The Annals of Applied Statistics*, 8(3): 1640–1670. MR3271347. doi: <http://dx.doi.org/10.1214/14-A0AS745>. 729
- Zidek, J. V., Sun, W., and Le, N. D. (2000). “Designing and integrating composite networks for monitoring multivariate Gaussian pollution fields.” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 49(1): 63–79. MR1817875. doi: <http://dx.doi.org/10.1111/1467-9876.00179>. 711

Acknowledgments

This paper is based on the doctoral thesis of the first author under the supervision of the second author. The authors thank the referees for numerous comments and suggestions that led to substantial improvement of the presentation. The second author wishes to thank the support from CNPq-BRAZIL.

Rejoinder*

Gustavo da Silva Ferreira[†] and Dani Gamerman[‡]

1 Introduction

We would like to start by thanking the Editor of BA for the opportunity to discuss our work and the discussants Peter J. Diggle, Michael G. Chipeta, James V. Zidek, Noel Cressie and Raymond L. Chambers for a very thorough evaluation of our contribution and for their thoughts on the topic. Our rejoinder to the discussion will be presented in the following topics: Preferential Sampling; Auxiliary Information; Models for $[X|S]$; Utility Functions; Sequential Design; and Approximations.

2 Preferential Sampling

Preferential sampling plays an important role in surveys routinely carried out by Official Statistics agencies, as those developed in the Brazilian Institute of Geography and Statistics (IBGE), the institution that one of us is affiliated to. So we are well aware of the relevance of this issue. In addition to the areas of application of preferential sampling cited by the discussants, it is important to mention the very topical area of publication bias (see Bayarri and DeGroot, 1993; Franco et al., 2014).

The link between preferential sampling and the methodology of survey-sampling presented by discussants is also very helpful in clarifying the similarities of approaches. We agree with Cressie and Chambers (hereafter CC) that papers from the latter may bring important aspects of sampling design to the context of Geostatistics. However, an important distinction between the approaches is needed. While in the context of survey methods the population size is generally fixed at a finite N , this feature is not true in the context of Geostatistics. In fact, in Geostatistics a finite value of N may only be associated with the discretization of a continuous process. Thus, part of the similarity between the approaches stems from the current limitation of many approaches to handle inference and prediction in Geostatistics appropriately due to the use of discrete approximations.

We agree with Chipeta and Diggle (hereafter CD) that preferential sampling is a method of adaptive design, which may depend on the previous design without relying on the underlying process S . Similarly, inference may be simplified after assuming that the process X is governed by the values of a spatially distributed covariate W , thus rendering conditional independence between X and S given W . The main difficulty is finding and quantifying such a covariate. We will return to this issue in the next section.

*Main article DOI: [10.1214/15-BA944](https://doi.org/10.1214/15-BA944).

[†]National School of Statistical Sciences, Brazilian Institute of Geography and Statistics, Rio de Janeiro, Brazil, gustavo.ferreira@ibge.gov.br

[‡]Department of Statistical Methods, Mathematical Institute, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil, dani@im.ufrj.br

3 Auxiliary Information

All discussants brought in the issue of replacement of some of the latent and unobserved random processes by observed quantities. These processes are a mere artifact to best represent our lack of knowledge about the true mechanism underlying the observation processes, if such truth exists. If we knew what variables cause our observational processes to behave as they do and were able to measure these quantities, we should use them. Otherwise, models that incorporate relevant features such as smoothness of these processes are a useful alternative. This issue is not only relevant to Spatial Statistics but to any other area of Statistics where absence of complete knowledge of the sources of variation forces the use of qualitative information in the form of latent processes, and Gaussian processes are one of the most adequate choices as a first step.

We agree that covariates could (and should) be part of all model components. Covariates may also be used to construct a deterministic intensity function of the point process X , or as proxy of a random process S in a model with a random intensity function. They may reduce remaining spatial heterogeneity in the mean or covariance structure of S , when available. CC suggested the use of other covariates Z , assumed to be highly correlated with the process S , as a proxy. This is a practical solution that needs to be used carefully. Despite simplifications produced in inference, it is important to bear in mind that this solution may also introduce another source of error in the model.

However, despite being a natural choice, we do not always have available covariates. In other situations, the covariates are available only in part of the region of interest. In these cases, it is often necessary to interpolate values before including them in the model, adding more uncertainty to the results.

4 Models for $[X|S]$

We agree that an appropriate model is crucial and robustness considerations are even more important here. Many discussants of Diggle et al. (2010) seemed to agree that, in practical situations, it is unlikely that the design is governed by a log-Gaussian Cox process. This issue is in line with our views expressed in the first paragraph of the previous section. In some cases, it may be enough to assume that the intensity of the point process is somehow proportional to the underlying process at an appropriate scale. In these cases, a log-linear intensity function may be seen as the first option for approximation and may thus be a good starting point to mitigate and understand the consequences of a preferential sample.

In addition to methods for obtaining a robust design against selection bias, alternative specifications for $[X|S]$ may also contribute to a satisfactory model-based inference. All discussants have correctly expressed concern about the sensitivity of the inference to the choice of the model for the intensity function. We concur with their concern and particularly welcome the use of non-parametric specifications to replace the parametric specification of our paper. This is bound to lead to a more flexible and hence robust alternative to the global linearity imposed by the $\alpha + \beta S$ predictor over the entire region of interest.

One approach following this path allows the regression coefficients' processes α and β to vary locally over space. Inference for these models may be performed in a simplified fashion, via discretization of coefficients over sub-regions (Pinto Junior et al., 2015), or exactly, by retaining the continuous variation of the infinite dimensional regression coefficient process (Gonçalves and Gamerman, 2015). Another non-parametric approach was proposed by Kottas and Sansó (2007), based on mixtures of Dirichlet process priors. Their idea could be adapted to a prior for the intensity of a log-Gaussian Cox process that (instead of being concentrated on) is only centered at $\alpha + \beta S$, again allowing more flexibility for the intensity.

Zidek also suggested an interesting situation where the generating process for the locations is based on another process S^* . In this case, it is first necessary to assess the degree of dependence between these two processes, that is, to evaluate if $[S, S^*] = [S][S^*]$. If so, the researcher can consider the points X as ancillary for S and proceed with the non-preferential, standard inference as usual. Otherwise, it may be necessary to establish some form of dependence for $[S, S^*]$ to be able to proceed with inference. A bivariate specification for $[S, S^*]$ may be one way forward (see Gamerman et al., 2007; Crainiceanu et al., 2008).

In conclusion, our model for $[X|S]$ seems to be able to highlight the effects of preferential sampling in the absence of a better understanding of the true causes of variation or of relevant covariates, but more flexible forms are needed.

5 Utility Functions

We agree with CC's suggestions about notation to improve the paper understanding. In particular, we consider pertinent the suggestion of making s_d explicit in the utility function.

The choice of a particular utility function is a crucial step to obtaining the optimal design. For this reason, we completely agree with CC that the questions *how much?* and *why?* cannot be set aside while the researcher is planning a new location sample.

Utility functions based on predictive variance reductions have long been recognized as appropriate to measure improvements in predictive accuracy. Zidek noted that more general evaluation (of predictive/estimation performance) may require the use of other utility functions. We already presented at least one alternative formulation early in the paper to emphasize our adherence to this point and to detach ourselves from the initial approaches based only on a single function, evaluating predictive errors. Our methods apply equally well to any quantifiable utility function.

The combination of different goals — e.g., reducing predictive errors, reducing uncertainty with respect to S , identifying thresholds, reducing estimation errors and evaluating costs involved — allows the researcher to obtain designs in complex situations. Examples of complex utility functions with competing goals can be found in the recent work of Müller et al. (2004), Ruiz-Cárdenas et al. (2012) and Ferreira (2015) to name a few. Note that these goals may be competing, e.g., reduction of the nugget effect estimation error assigns more utility to regions close to locations already sampled

whereas reduction of predictive error assigns more utility for regions far from the points already sampled. The additional challenge in this case is to weigh the various objectives involved.

Alternatively, the use of entropy in the utility function is usually recommended for situations where multiple goals are involved (Caselton and Zidek, 1984) but this is not an easily implementable solution in practice. We do believe that in practical situations the utility function needs to be specified by whoever is in charge of the analysis, be it a regulatory body, a government institution or a researcher, with the help of a statistician. The responsible entity must be able to value the worth of the information each component of the utility provides. If one can answer, for example, why predictive variance reduction is relevant, one must be able to assign its monetary value; otherwise, one must reassess the relevance of each component included in the utility function. This value is more easily combined with readily quantifiable monetary components such as cost associated with each new location. An economist specialized in the area of the study may be a useful addition to the team setting up this enterprise at this stage.

6 Sequential Design

The approach we used for design (Müller, 1999) allows the planning of m new sampling sites according to the utility function defined by the researcher. This design plan can be sequential or not, while the underlying process S can be static or dynamic in time. In cases where S is a dynamic process, planning of a sequential design scheme is a natural choice. However, in the case where S is fixed in time, one can also plan a sequential design, although it is not possible to ensure that the resulting design will be optimal. Actually, a sequential sample design can be a simple and cheap solution, especially in situations where the costs are not fixed and can increase before obtaining the desired new sample locations.

We anticipate difficulties to assigning a distribution for $[d|S, \mathbf{x}]$ without reducing the flexibility of the model, when a sequential-sampling-design strategy to update θ and S through $[d|\mathbf{x}, \mathbf{y}]$ is built. It would be necessary to assess whether the initial sample is preferential in cases suggested by CC with a pilot study. If it is possible to assume that the pilot phase is non-preferential, then this information will be ancillary to inference. On the other hand, if the pilot sample is preferential, then it is possible to incorporate this information by performing modifications in $[X|S]$. Note that this may remove the Poissonity of the model for $[X|S]$, due to repulsions that may occur around previously selected locations. This will imply more difficulty for likelihood approaches and opens up for interesting research questions.

The case of a dynamic underlying process is more complex. We visualize the following generalization of the 2-stage set-up proposed by CD. In cases where samples are taken at different times in a multi-stage scenario, we may have

$$[Y, X, S] = \prod_t [Y_t | Y_{1:t-1}, X_{1:t}, S_t] [X_t | X_{1:t-1}, S_{1:t}] [S_t | S_{1:t-1}],$$

where $Z_{a:b} = (Z_a, Z_{a+1}, \dots, Z_{b-1}, Z_b)$, for $a < b$ integers, and dependence on hyperparameters was removed from the notation as in CD. Conditional independence between observations given the corresponding underlying processes may be assumed in some cases leading to $[Y_t | Y_{1:t-1}, X_{1:t}, S_t] = [Y_t | X_t, S_t]$. Similarly, $[X_t | X_{1:t-1}, S_{1:t}] = [X_t | S_t]$ may be assumed for some sampling schemes, although the general formulation may be required in some cases (see the paragraph above). Further simplification such as those suggested by CD may be assumed, depending on the situation. Ferreira (2015) worked on a similar structure, simplified by his non-preferential sampling scheme.

In the practical situation presented by CD, where the sampling order is a crucial factor, a simple utility function based on predictive variance reductions or exceedances probably would not be enough to produce a satisfactory sample design. In challenging situations like this, it would be necessary to choose more complex forms to reward each sample unit, at each time, in a sequential sampling scheme.

7 Approximations

Questioning the stationarity assumption is a mandatory task in any Spatial Statistics problem. Stationarity can always be seen as an approximation to a more complex underlying dependence. But it turns out that in many practical situations it has proved to be a reasonable, viable solution. Alternatively, approaches based on convolution process (Higdon, 2002) or the use of more flexible (non-parametric) structures to enhance the spatial dependence structure of S can be used. Again, covariates are always relevant options to handle the large-scale heterogeneity considered by CD. Obviously, an increase in the complexity of the model can also complicate the evaluation of the utility function used to obtain the optimal design and parsimony may have to be called into action.

We agree with Zidek that other approximation of $V(S|\mathbf{x}, \mathbf{y})$ could be used in this step. A simpler, but expensive, alternative is to generate sub-chains in order to estimate this quantity during MCMC. Alternative analytical approaches may be preferred to avoid the increasing computational cost. It is important to emphasize that

- The approximation was only used to evaluate the utility function, since an analytical expression for $V(S|\mathbf{x}, \mathbf{y})$ is not available; the values of S were always sampled from the correct distribution $[S|\mathbf{y}, \mathbf{x}]$;
- Other utility functions, as those used in our case study, may not require any approximation.

References

- Bayarri, M. J. and DeGroot, M. H. (1993). “The analysis of published significant results.” In: *Rassegna di Metodi Statistici ed Applicazioni* (W. Racugno, ed.), 19–41. Pitagora. [MR1223386](#). [753](#)
- Caselton, W. F. and Zidek, J. V. (1984). “Optimal monitoring network designs.” *Statistics & Probability Letters*, 2(4): 223–227. [756](#)

- Crainiceanu, C. M., Diggle, P. J., and Rowlingson, B. (2008). “Bivariate binomial spatial modeling of *Loa loa* prevalence in tropical Africa.” *Journal of the American Statistical Association*, 103(481): 21–37. MR2420211. doi: <http://dx.doi.org/10.1198/016214507000001409>. 755
- Diggle, P. J., Menezes, R., and Su, T.-L. (2010). “Geostatistical inference under preferential sampling (with discussion).” *Applied Statistics*, 59(2): 191–232. MR2744471. doi: <http://dx.doi.org/10.1111/j.1467-9876.2009.00701.x>. 754
- Ferreira, M. A. (2015). “Inhomogeneous evolutionary MCMC for Bayesian optimal sequential environmental monitoring.” *Environmental and Ecological Statistics*, 1–20. 755, 757
- Franco, A., Malhotra, N., and Simonovits, G. (2014). “Publication bias in the social sciences: Unlocking the file drawer.” *Science*, 345(6203): 1502–1505. 753
- Gamerman, D., Salazar, E., and Reis, E. (2007). “Dynamic Gaussian process priors, with applications to the analysis of space-time data (with discussion).” In: *Bayesian Statistics* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, eds.), volume 8, 149–174. Oxford University Press. MR2433192. 755
- Gonçalves, F. and Gamerman, D. (2015). “Exact Bayesian inference in spatio-temporal Cox processes driven by multivariate Gaussian processes.” Technical report, Statistical Methods Laboratory, Federal University of the Rio de Janeiro. 755
- Higdon, D. (2002). “Space and space-time modeling using process convolutions.” In: *Quantitative Methods for Current Environmental Issues* (C. W. Anderson, V. Barnett, P. C. Chatwin and A. H. El-Shaarawi, eds.), 37–56. Springer. MR2059819. 757
- Kottas, A. and Sansó, B. (2007). “Bayesian mixture modeling for spatial Poisson process intensities, with applications to extreme value analysis.” *Journal of Statistical Planning and Inference*, 137(10): 3151–3163. MR2365118. doi: <http://dx.doi.org/10.1016/j.jspi.2006.05.022>. 755
- Müller, P. (1999). “Simulation-based optimal design.” In: *Bayesian Statistics* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.), volume 6, 459–474. Oxford University Press. MR1723509. 756
- Müller, P., Sansó, B., and De Iorio, M. (2004). “Optimal Bayesian design by inhomogeneous Markov chain simulation.” *Journal of the American Statistical Association*, 99(467): 788–798. MR2090911. doi: <http://dx.doi.org/10.1198/016214504000001123>. 755
- Pinto Junior, J. A., Gamerman, D., Paez, M. S., and Fonseca Alves, R. H. (2015). “Point pattern analysis with spatially varying covariate effects, applied to the study of cerebrovascular deaths.” *Statistics in Medicine*, 34(7): 1214–1226. 755
- Ruiz-Cárdenas, R., Ferreira, M. A., and Schmidt, A. M. (2012). “Evolutionary Markov chain Monte Carlo algorithms for optimal monitoring network designs.” *Statistical Methodology*, 9(1): 185–194. MR2863607. doi: <http://dx.doi.org/10.1016/j.stamet.2011.01.009>. 755